Nikita Markov

nikitamarkov.work@gmail.com | desire32.github.io/blog | github.com/Desire32 | Linkedin

Profile

Software Engineer and research-oriented developer working at the intersection of machine learning, software systems, and next-generation communication technologies.

Skills

Programming: Python (3.12+), C++, Go (basic)
ML & Deep Learning: PyTorch, Transformers (HF), scikit-learn, NumPy, Pandas
LLMs & NLP: LangChain, LoRA, RAG, ASR
Model Optimization & Inference: ONNX, Apache TVM, TensorRT
Data & Infrastructure: PostgreSQL, MongoDB, Redis, Docker, Git, AWS (EC2, RDS)
MLOps & Monitoring: MLflow, Weights & Biases

Education

University of Central Lancashire *Cyprus*
BSc in Computer Engineering / Computing *Grade: First Class | 09/22 – 09/26*
*Thesis (**Research project**): **Performance Evaluation of UE-VBS as Computational and Storage Hub (CSHs) in 6G Networks**.*
Design, implement and evaluate a prototype *UE-VBS (User Equipment Virtual Base Station)* communication system to demonstrate ***low-latency edge processing***. reducing latency and cost in future *6G RAN* systems.

Experience

RIF Internship — Abasis AI *Jul 2025 – Aug 2025*
Cypriot ASR dialect model, News post

- Developed Cypriot Greek ASR system **in 6 weeks** under tight resource constraints using **Wav2Vec2** architecture, achieving performance through fine-tuning on **90,000+ audio-text pairs** and custom dataset creation from parliamentary recordings and news broadcasts
- Implemented **KenLM n-gram language modeling** to enhance transcription accuracy, reducing Word Error Rate (**WER**) ***by 7 percents*** by integrating a 6-gram model trained on 89,000+ Cypriot dialect text pairs from multiple sources
- Built end-to-end ML pipeline including **MLflow** experiment tracking, custom data cleaning and comparative evaluation of 8+ ASR systems using ***WER/CER*** metrics

InSPIRE Research Center - Research Assistant *Oct 2024 - Jan 2026*
[*IEEE COMPSAC***] Quantization at the Edge: Evaluating Inference Performance and Quality for SLM Integration in Virtual Worlds** — *UNDER REVIEW*

* Led optimization of LLM inference stack, achieving **2.3× model size reduction** (1.1B → 470M) via **TVM** + **MLC-LLM** quantization while preserving output quality.
* Architected a production-grade LLM runtime platform with dynamic model embedding routing, **multi-scheme quantization (INT3–INT8)**, RAG-based knowledge retrieval, and full-stack runtime observability.
* Built a Dockerized performance benchmarking framework using **ONNX** and TensorRT (**trtexec**) for NVIDIA Jetson deployment.

[*IEEE COMPSAC***] Enhancing Digital Heritage Experiences: Evaluating Fine-Tuned LLM Integration** — Publication

- A modular fine-tuning pipeline for various architectures with ***qLoRA***, a dual-channel architecture with semantic search and ***Word2Vec*** for optimizing data access, as well as ***MLflow*** integration for tracking and analyzing model performance.

[*IEEE COMPSAC***] Developing a Cyber-Physical-Social Metaverse System for Cultural Experiences** — Publication

- Accepted into a peer-reviewed research paper with a competitive **22%** ***acceptance rate***.
- A chatbot system based on ***LangChain*** with local integration of LLM Mistral-7B via Ollama and other ***Upstash Redis*** memory stores, a scalable ***Flask-based REST API*** for session management and streaming responses, and an extension to ***AWS EC2*** for security and production.

Languages

English (C1), Russian (Native)